AL/CF-TR-1995-0155

PITCH AND TIME-SCALE MODIFICATION OF SPEECH: A REVIEW OF THE LITERATURE

RAYMOND E. SLYH

CREW SYSTEMS DIRECTORATE
BIODYNAMICS AND BIOCOMMUNICATIONS DIVISION
WRIGHT-PATTERSON AFB OH 45433-7901

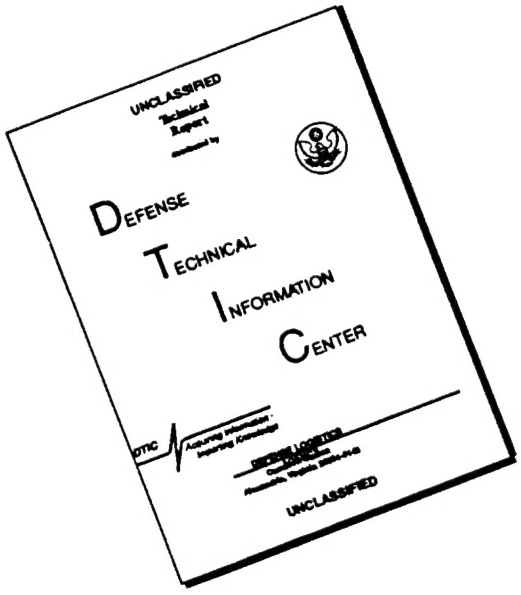
AUGUST 1995

INTERIM REPORT FOR THE PERIOD MAY 1994 TO MAY 1995

AIR FORCE MATERIEL COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO 45433-6573

ARMSTRONG LABORATORY

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

NOTICE

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner, licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

Please do not request copies of this report from the Armstrong Laboratory. Additional copies may be purchased from:

National Technical Information Service
5285 Port Royal Road
Springfield VA 22161

Federal Government agencies and their contractors registered with Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center
Cameron Station
Alexandria VA 22314

TECHNICAL REVIEW AND APPROVAL

AL/CF-TR-1995-0155

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

THOMAS J. MOORE, Chief
Biodynamics and Biocommunications Division
Crew Systems Directorate
Armstrong Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 1995	Interim - May 1994 to May 1995

4. TITLE AND SUBTITLE
Pitch and Time-Scale Modification of Speech: A Review of the Literature

5. FUNDING NUMBERS
PE - 62202F
PR - 7231
TA - 21
WU - 04

6. AUTHOR(S)
Raymond E. Slyh

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Armstrong Laboratory, Crew Systems Directorate
Biodynamics and Biocommunications Division
Human Systems Center
Air Force Materiel Command
Wright-Patterson AFB OH 45433-7901

8. PERFORMING ORGANIZATION REPORT NUMBER
AL/CF-TR-1995-0155

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This report summarizes the literature in the areas of pitch and time-scale modification of speech. Based on the discussions in the literature, the report recommends the following four techniques for further evaluation: (1) the pitch-synchronous overlap-add technique; (2) the sinusoidal analysis/synthesis system of Quatieri and McAulay; (3) the harmonic plus noise model of Laroche, Stylianou, and Moulines; and (4) the time-domain harmonic scaling technique. All four techniques are capable of modifying both the pitch and the time scale of speech signals.

14. SUBJECT TERMS	15. NUMBER OF PAGES
pitch modification, time-scale modification, speech processing	36
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

This page intentionally left blank.

PREFACE

This report was prepared in the Biodynamics and Biocommunications Division, Crew Systems Directorate, of the Armstrong Laboratory, Wright-Patterson Air Force Base, Ohio. The work was performed under Project 7231, Biomechanics of Air Force Operations, Task 723121, Biocommunications, Work Unit 72312104, Bioacoustics and Biocommunications Research. The author wishes to thank both Dr. Timothy Anderson of Armstrong Laboratory and Ms. Marty Luka of Systems Research Laboratories for reviewing drafts of this report.

TABLE OF CONTENTS

LIST OF FIGURES

This page intentionally left blank.

INTRODUCTION

This report summarizes the literature in the areas of pitch and time-scale modification of speech. The report primarily focuses on the pitch modification of speech. However, since some pitch modification techniques simultaneously modify the time scale of a speech signal (and vice versa), the report also briefly covers the time-scale modification literature. Based on the claims made in the literature, this report recommends four techniques for further consideration—namely, the pitch-synchronous overlap-add technique; the sinusoidal analysis/synthesis system of Quatieri and McAulay; the harmonic plus noise model of Laroche, Stylianou, and Moulines; and the time-domain harmonic scaling technique. All four techniques are capable of modifying both the pitch and the time scale of speech signals.

The problems of pitch modification and time-scale modification are perhaps best discussed in the context of a simple source/filter model of speech signals. In such a model, the speech signal is considered to be the output of a linear time-varying filter excited by a time-varying source. The time-varying source consists of two major components in changing proportions—namely, a quasiperiodic pulse train and a noise-like portion. When the source consists of mostly the pulse train component, the output speech is quasiperiodic and considered to be "voiced." When the source consists of mostly the noise-like component, the output speech is noise-like and considered to be "unvoiced."

In general, a segment of speech has mixed voicing; in other words, the source consists of both a quasiperiodic pulse train portion and a noise-like portion. For segments with mixed voicing, the filter applied to the noise-like portion of the source may be different from the filter applied to the quasiperiodic pulse train. Thus, a speech signal, $s(t)$, can be written as

$$s(t) = \int_{-\infty}^{\infty} h_t^V(\tau) u^V(t-\tau)\, d\tau + \int_{-\infty}^{\infty} h_t^U(\tau) u^U(t-\tau)\, d\tau, \tag{1}$$

where

- $u^V(t)$ is the voiced portion of the source (*i.e.*, the quasiperiodic pulse train),
- $h_t^V(t)$ is the linear time-varying filter that acts on the voiced portion of the source,
- $u^U(t)$ is the unvoiced portion of the source (*i.e.*, the noise-like portion of the source), and
- $h_t^U(t)$ is the linear time-varying filter that acts on the unvoiced portion of the source.

The pitch modification problem can be viewed in both a narrow sense and a wide sense. In a narrow sense, pitch modification consists of modifying the time-varying fundamental

frequency of $u^V(t)$ while making no other changes to $u^V(t)$. In a wider sense, pitch modification consists of modifying the time-varying fundamental frequency while also judiciously modifying the time-varying spectral envelope of $u^V(t)$.

In general, the speech model of Equation 1 is not unique. The spectral shape of $s(t)$ is partially due to the sources and partially due to the filters, and the contributions of either the sources or the filters to the overall spectral shape of the signal can be modified as long as other factors are adjusted to compensate for these modifications. In particular, one can modify the spectral shape of $u^V(t)$ and still obtain the same $s(t)$ provided that one modifies the spectral shape of $h_t^V(t)$ properly. In the literature, various different shapes have been tried for $u^V(t)$, ranging from a simple impulse train as used in standard linear predictive coding to more complex pulse trains representing the glottal flow. An example of the latter is shown in Figure 1(a).

When $u^V(t)$ is a simple impulse train, pitch modification consists of simply changing the spacings of the impulses. However, when the pulse shapes in $u^V(t)$ are more complex, modifications of the pulse shape (and hence of the spectral envelope of $u^V(t)$) may also be necessary. For example, Figure 1(b) shows that moving pulses close together without modifying the shape of the pulses can eliminate the closed phase of the glottal flow. (The closed phase in each period of the waveform is the portion with zero magnitude.) On the other hand, Figure 1(c) shows a number of pulses of the same fundamental frequency as those in Figure 1(b), but with the pulse shape compressed. In this case, a closed-phase portion exists for each period of the waveform. The point of this example is that the wider view of pitch modification as consisting of both modifications to the time-varying fundamental frequency and modifications to the the time-varying spectral envelope of $u^V(t)$ has certain merits. This report considers the pitch modification problem in the wider sense.

The various pitch modification techniques in the literature fall into two main categories—namely, frequency-scaling techniques and frequency-resampling techniques. Frequency-scaling techniques modify both the fundamental frequency and the spectral envelope of $u^V(t)$, while frequency-resampling techniques attempt to modify only the fundamental frequency of $u^V(t)$. The following definitions serve to delineate the two categories.

> **Frequency Resampling:** The process of changing the fundamental frequency of a quasiperiodic segment of a signal without modifying the segment's time duration or its spectral envelope.

> **Frequency Scaling:** The process of compressing or expanding the spectrum of a signal without modifying the time duration of the signal. In other words, frequency-scale modification is the process of compressing or expanding the short-time Fourier transform (STFT) of a speech signal only along the frequency axis.

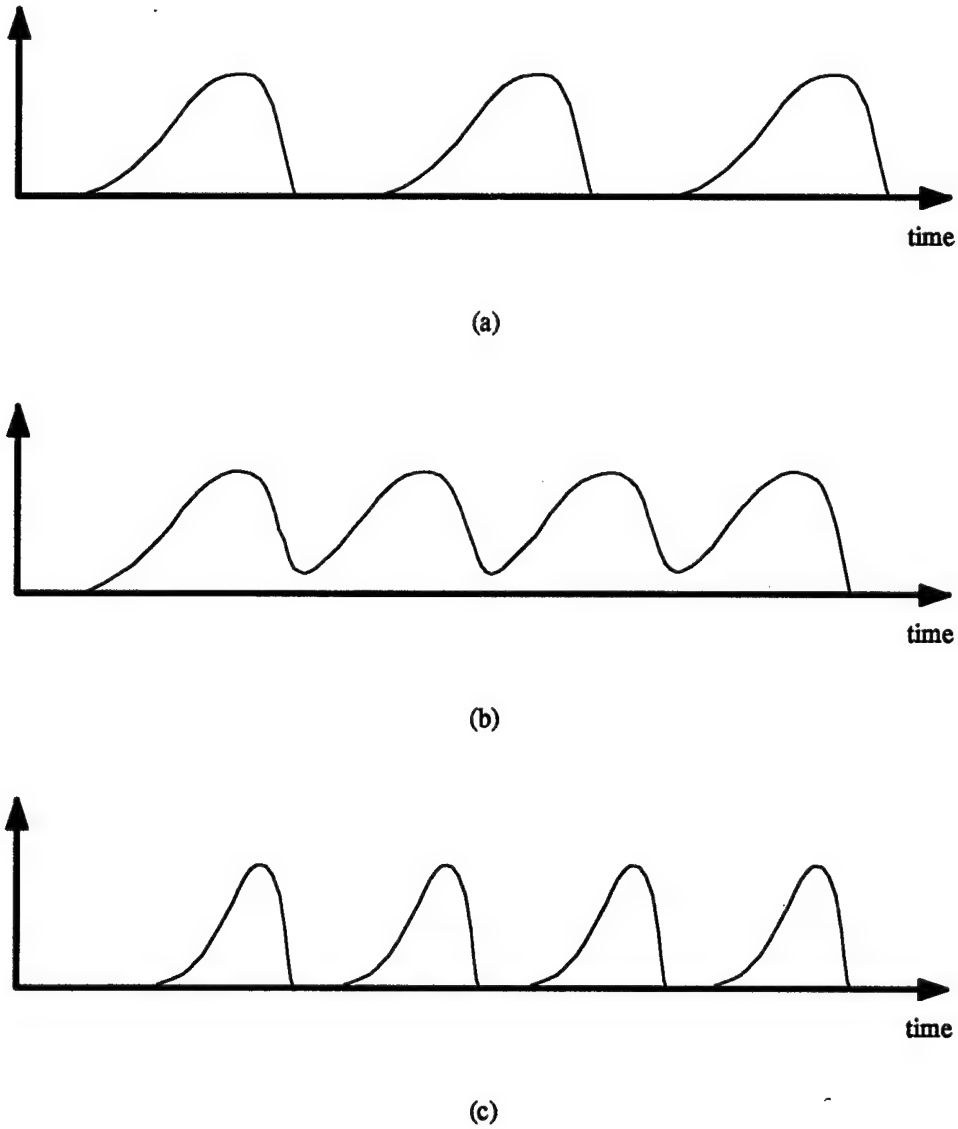


Figure 1: Source Waveforms Showing (a) Plot of the Glottal Flow for a Voiced Phoneme, (b) Pitch Modification of the Original Glottal Flow Waveform Accomplished by Overlapping and Adding the Individual Pulses, and (c) Pitch Modification of the Original Glottal Flow Waveform Accomplished by Compressing the Pulses in Time

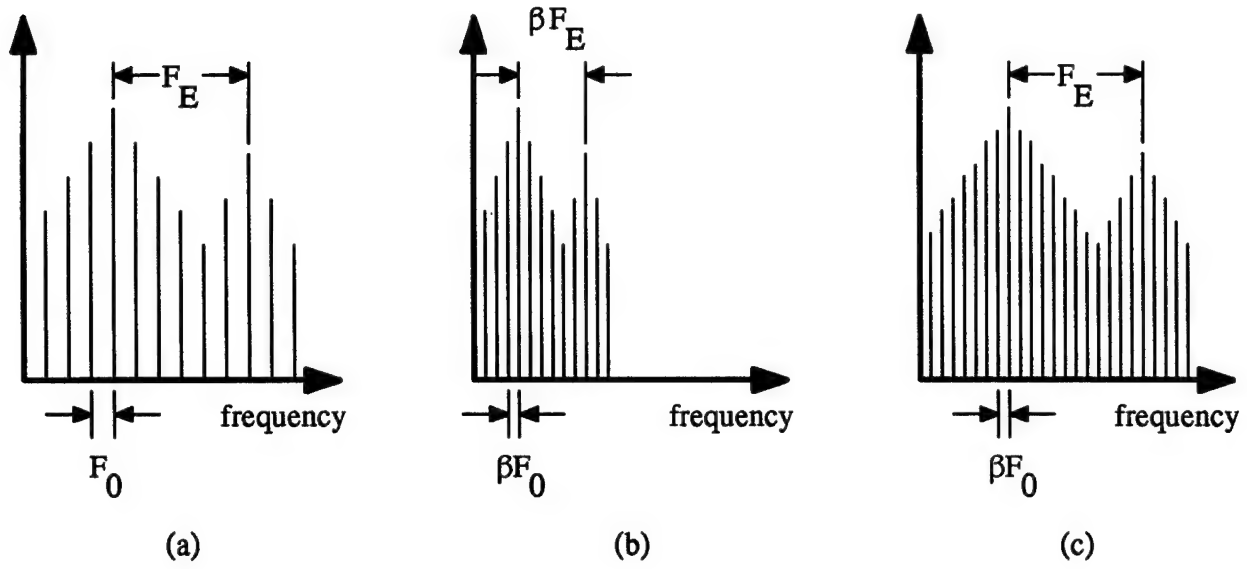


Figure 2: Spectra of (a) an Unmodified Signal, (b) a Signal Frequency-Scaled by a Factor of β, and (c) a Signal After Application of Frequency Resampling by a Factor of β

Note that it is not clear how closely the spectral envelope modifications of frequency scaling resemble those required to properly transform $u^V(t)$.

Figure 2 illustrates the difference between frequency resampling and frequency scaling. Figure 2(a) shows the spectrum of an unmodified periodic signal. The spectral envelope of the signal contains two peaks with a frequency spacing of F_E. The fundamental frequency is F_0. Scaling the frequency axis by a factor of β results in the modified spectrum shown in Figure 2(b). Here, the spacing between the two peaks of the spectral envelope is βF_E, and the modified fundamental frequency is βF_0. Thus, frequency scaling modifies both the harmonic structure and the spectral envelope of the signal. Resampling the frequency axis by a factor of β results in the modified spectrum shown in Figure 2(c). The modified fundamental frequency is again βF_0, but the spectral envelope is the same as that of the original signal. In particular, the spacing between the two peaks in the spectral envelope is the original spacing of F_E.

For completeness, a definition of the time-scale modification process is as follows.

> **Time-Scale Modification:** The process of compressing or expanding the time duration of a signal without modifying the apparent local frequency content of the signal. In other words, time-scale modification is the process of compressing or expanding the STFT of a speech signal only along the time axis.

MAJOR PITCH MODIFICATION TECHNIQUES

This section presents the major pitch modification techniques: techniques based on simple linear prediction models for speech; the multiband excitation vocoder; the time-domain harmonic scaling method and linear prediction; the sinusoidal analysis/synthesis system of Quatieri and McAulay; the harmonic plus noise model of Laroche, Stylianou, and Moulines; and the pitch-synchronous overlap-add method.

Basic Linear Predictive Coding Models of Speech

Basic linear predictive coding (LPC) models of speech provide some of the most straightforward ways for modifying the pitch of a speech signal [1–5]. Figure 3 shows an LPC-based pitch modification method. First, the system performs LPC analysis on short segments of the original signal resulting in a set of filter coefficients for each segment and a residual signal. (Filtering the residual signal with the time-varying filter coefficients returns the original signal.) Second, the system forms a synthetic excitation signal, where the excitation signal consists of white noise for the unvoiced segments and periodic impulse trains for the voiced segments. The spacing between the impulses, T_0, varies according to the desired pitch, $T_0 = \frac{1}{F_0}$. Third, the system forms the pitch-modified signal by filtering the synthetic excitation signal with the time-varying linear filter formed by the LPC coefficients.

Although this method is conceptually simple and easy to implement, the output speech has a distinct synthetic quality due to the simplified excitation model [1, 3–5]. This fact has prompted researchers to develop excitation models that produce more natural sounding speech. In [3], Milenkovic models the voiced-speech excitation as shown in Figure 4. Pitch modification using this excitation could be accomplished by varying the spacing between the pulses. In multipulse LPC models [6–8], the excitation consists of a small number of impulses (generally eight to 10) over each 10 msec frame. In [9], Caspers and Atal investigated two different methods for modifying the pitch of multipulse LPC speech. The first method modified the length of individual pitch periods by linearly scaling the time axis of the multipulse excitation. The second method modified the length of individual pitch periods by adding or subtracting zeros in the excitation. Both methods modified the length of the excitation, so pitch periods were added or subtracted from the excitation in order to obtain an excitation of the same length as the original excitation. Caspers and Atal found that the second pitch modification method (*i.e.*, the addition or subtraction of zeros in the excitation) introduced very little distortion, while the linear scaling of the time axis produced significantly more distortion.

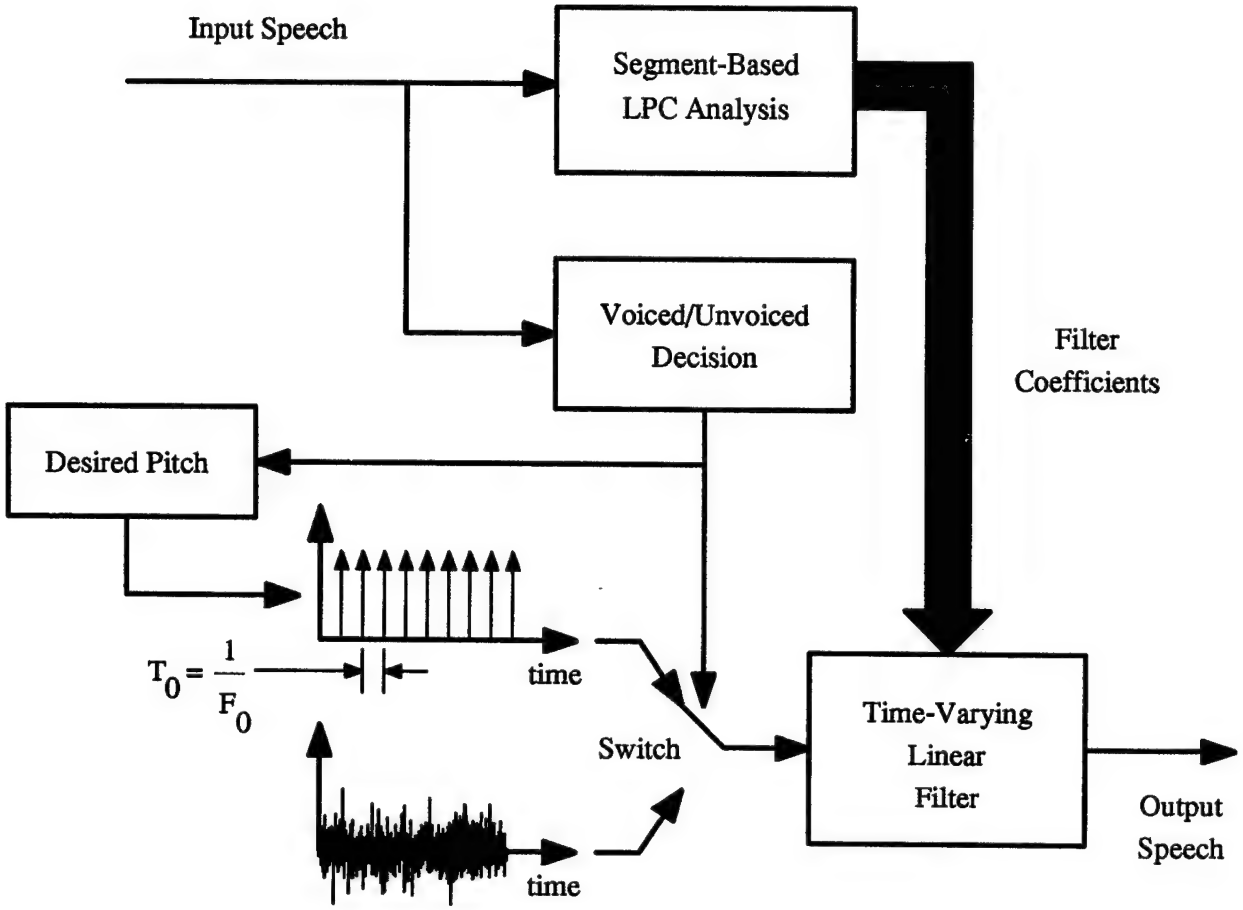


Figure 3: Pitch Modification Based on Linear Predictive Coding Analysis of the Speech Signal

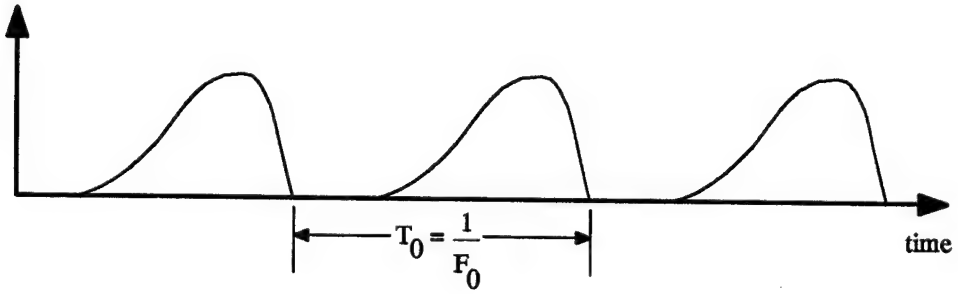


Figure 4: Sketch of the Voiced-Speech Excitation Signal Used by Milenkovic (1993)

The Multiband Excitation Vocoder

As noted in [1,2,10], voiced-speech segments generally contain both harmonic and inharmonic components, and modeling voiced-speech segments as purely harmonic signals leads to synthetic speech with a buzzy quality. For these reasons, Griffin and Lim developed the multiband excitation (MBE) vocoder [1,2,10]. The MBE vocoder divides the spectrum of the excitation into a number of adjacent frequency bands with each band labeled as voiced or unvoiced. Thus, the voiced-speech excitation model for the MBE vocoder contains both voiced and unvoiced components.

Figure 5 shows a diagram of a pitch modification system based on the MBE vocoder. Over small speech segments, the system estimates the spectral envelope and the pitch of the speech and makes voiced/unvoiced decisions about the speech. The system divides the spectrum of the speech into several adjacent frequency bands, and makes a voiced/unvoiced decision for each band. These voiced/unvoiced decisions yield a voicing indicator function in the frequency domain, where the value is one for voiced bands and zero for unvoiced bands. Using the desired pitch, the system creates the voiced portion of the excitation by including those harmonics of the desired fundamental frequency that fall within the voiced-speech regions as indicated by the voicing indicator function. The system forms the unvoiced portion of the excitation by multiplying the spectrum of a sample white noise sequence with a function indicating the unvoiced frequency bands (*i.e.*, a function that has a value of one for unvoiced frequency bands and a value of zero for voiced frequency bands). Finally, the system forms the output speech signal by scaling the excitation spectrum (the sum of the voiced and unvoiced portions) by the estimated spectral envelope.

Sampling-Rate Conversion and LPC

To improve the quality of speech from LPC-based systems, one can use the LPC residual signal as the excitation for the time-varying filter given by the LPC coefficients. One generally does not use the original residual as the excitation signal in coding applications, because the residual requires several bits to code. However, reducing the bit rate is of little concern in the pitch modification problem, so one can use the residual as the excitation to greatly improve the quality of the output speech.

Conversion of the sampling rate of the LPC residual of a speech signal is one way to change the pitch of the signal. Figure 6 shows a diagram of a pitch modification system based on modifying the LPC residual. Suppose that the LPC residual is originally sampled at a rate

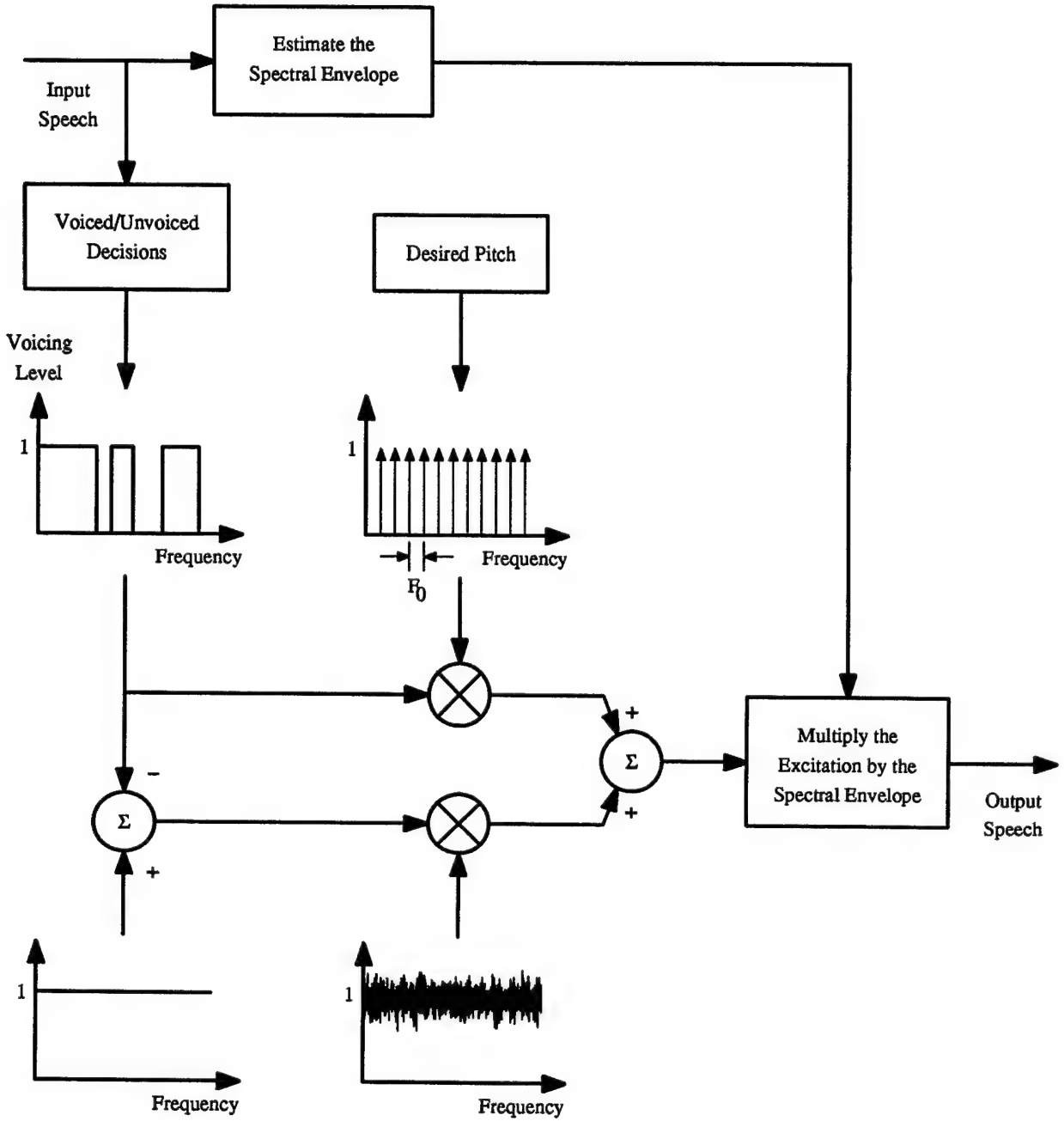


Figure 5: A Pitch Modification System Based on the Multiband Excitation Vocoder

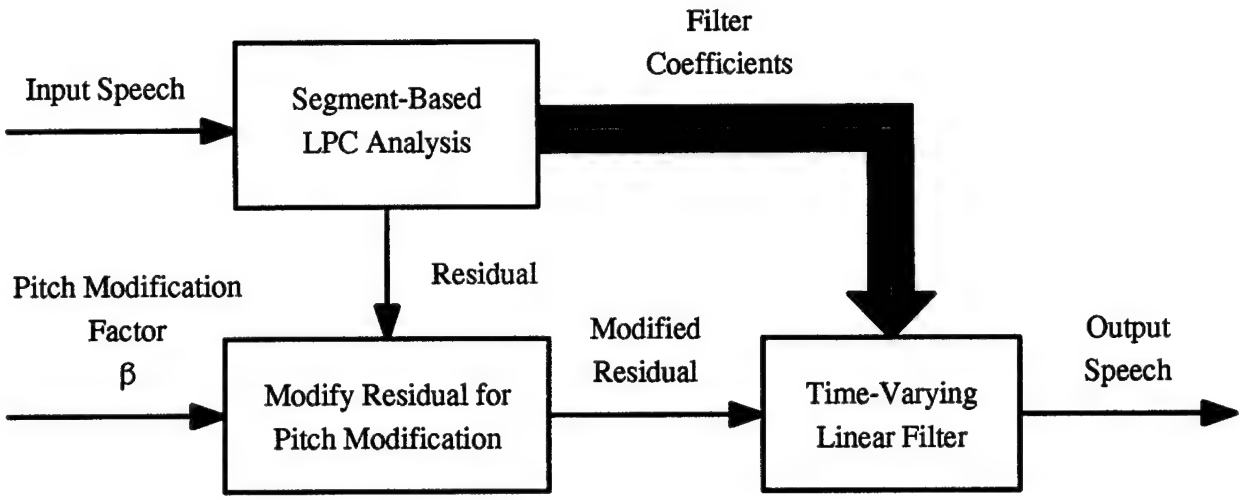


Figure 6: Diagram of a Pitch Modification System Based on Modifying the LPC Residual

of F_S in samples/second, and suppose that a system changes the sampling rate to F_N. If the system outputs points from the resampled residual at the new rate of F_N samples/second, then the output residual is a lowpass-filtered version of the original residual. However, if the system outputs points from the converted residual at the original sampling rate of F_S points/second, then the residual is scaled in frequency by the factor $\beta = \frac{F_S}{F_N}$. Note that this procedure changes the duration of the signal, so a time-scale modification technique must be used in conjunction with this technique. Methods for sampling-rate conversion can be found in [11–19].

There are two main drawbacks to using sampling-rate conversion as a means for pitch modification. The first drawback is that sampling-rate conversion scales the frequency response of the unvoiced portions of speech as well as the voiced portions. The second drawback is that fine time-varying pitch modification is difficult to perform using this technique.

Time-Domain Harmonic Scaling and LPC

Another pitch modification method based on modifying the LPC residual is the technique of time-domain harmonic scaling (TDHS) [20–27]. TDHS is a general technique for frequency scaling a signal (not just the LPC residual). Techniques similar to TDHS have been developed in [28, 29].

The TDHS-based pitch modification system has the same functional form as shown in Figure 6. The processing for TDHS depends on whether one wants to expand or compress the frequency spectrum. The steps required to scale the spectrum by a factor of β are as follows:

- Find two relatively prime integers μ and δ such that $\beta \approx \frac{\mu}{\delta}$. Let $C = \frac{\delta}{\mu}$. For frequency compression, $\mu < \delta$, which implies that $C > 1$. For frequency expansion, $\delta < \mu$, which implies that $C < 1$.

- Find the pitch period, T_0, and calculate an integer, N_P, such that $T_0 \approx N_P T$, where T is the sampling period. N_P is the approximate number of signal samples in one pitch period.

- Calculate the following integer values: $N = \mu N_P$ and $N_C \approx \frac{\mu N_P}{|\delta - \mu|} = \frac{N_P}{|C-1|}$.

- Calculate $T_C = \frac{T N_P}{N_C}$.

- Define $\alpha_C(n) \triangleq \left\lfloor \frac{n|\delta-\mu|}{\mu N_P} \right\rfloor = \left\lfloor \frac{n}{N_C} \right\rfloor$, where n is some integer and $\lfloor \cdot \rfloor$ denotes the floor operator.

- Define $h_N(\cdot) \triangleq N h(\cdot)$, where $h(\cdot)$ is a sampled version of an analog prototype lowpass filter (usually chosen to be a triangular window).

- Calculate

$$y^{\frac{1}{C}}(nCT) = \begin{cases} \sum_{i=0}^{m-1} x\left(nT + \alpha_C(n) N_P T - i N_P T\right) h_N\left(i N_C T_C + (n \bmod N_C) T_C\right), & \beta < 1 \\ \sum_{i=1}^{m} x\left(nT - \alpha_C(n) N_P T - i N_P T\right) h_N\left(i N_C T_C - (n \bmod N_C) T_C\right), & \beta > 1 \end{cases}$$

where $y^{\frac{1}{C}}(\cdot)$ is the frequency-scaled residual, $x(\cdot)$ is the original residual, and m is a small integer (generally, $m = 2$). For $\beta < 1$, N_C output samples are computed for every $N_C + N_P$ input samples. For $\beta > 1$, N_C output samples are computed for every $N_C - N_P$ input samples.

The TDHS technique is capable of both pitch modification and time-scale modification. If samples of $y^{\frac{1}{C}}(nCT)$ are output at the rate $\frac{1}{CT}$, then the modified residual has a scaled bandwidth and the same duration as the original residual. If samples of $y^{\frac{1}{C}}(nCT)$ are output at the rate $\frac{1}{T}$, then the modified residual has the same bandwidth as the original residual but a modified time scale compared with the original residual. Note, however, that the pitch and time-scale modifications are often approximate due to the approximations noted in the above steps.

General STFT-Based Pitch Modification Techniques

There are many frequency scaling techniques based on the short-time Fourier transform [5, 30–44]. These techniques calculate the STFT of a signal, modify the amplitudes and

phases of the STFT, and calculate the inverse of the modified STFT to form the modified signal. We can apply these techniques to the linear prediction residual as in Figure 6 and often directly to the speech signal itself in order to modify the pitch of the speech signal. These techniques often generate reverberation and other artifacts [45]. The reverberation arises from improper phase modifications to the excitation spectrum, while some of the other artifacts arise from modifying the spectrum of the unvoiced portions of the speech signal. The next subsection outlines an STFT-based system that adjusts the phases in order to greatly reduce the degree of reverberation in the output signal.

The Sinusoidal Analysis/Synthesis System of Quatieri and McAulay

The sinusoidal analysis/synthesis (SAS) system of Quatieri and McAulay is a versatile system for affecting time-scale, frequency-scale, and pitch modifications on speech signals [45]. Quatieri and McAulay developed the system in a series of papers [45–51], while Serra and Smith independently developed a similar system in [52].

The SAS model of the speech signal is

$$s(t) = \sum_{j=1}^{N(t)} a_j(t) M(\omega_j(t), t) \cos\left[\omega_j(t)\,(t - t_{p_i}) + \psi\left(\omega_j(t), t\right)\right],$$

where

- $N(t)$ is the number of frequency components at time t,
- $\omega_j(t) = 2\pi f_j(t)$ is the jth angular-frequency component at time t,
- t_{p_i} is the most recent pitch onset time for time t,
- $a_j(t)$ is the amplitude of the jth frequency component of the excitation at time t,
- $M(\omega_j(t), t)$ is the magnitude of the vocal tract frequency response evaluated at the frequency $\omega_j(t)$ for time t, and
- $\psi\left(\omega_j(t), t\right)$ is the phase of the vocal tract response evaluated at the frequency $\omega_j(t)$ for time t.

Note that the pitch onset time is the time for which all of the excitation sinusoids add coherently prior to their modification by the vocal tract response. Also note that the pitch onset times relate to each other through the time-varying pitch period, T_0, as

$$t_{p_i} = t_{p_{i-1}} + T_0.$$

Let

$$A(\omega_j(t),t) \triangleq a_j(t)M(\omega_j(t),t), \tag{2}$$

$$\theta(\omega_j(t),t) \triangleq \omega_j(t)(t-t_{p_i}) + \psi(\omega_j(t),t), \tag{3}$$

then $s(t)$ is more compactly written as

$$s(t) = \sum_{j=1}^{N(t)} A(\omega_j(t),t)\cos\left[\theta(\omega_j(t),t)\right].$$

The SAS system determines the various components in Equations 2 and 3 as follows. From the original speech signal, the SAS system estimates the values of $A(\omega_j,t)$ and $\theta(\omega_j(t),t)$, as well as the fundamental frequency and pitch onset times. For each j, the system computes the various quantities in Equations 2 and 3 as

$$a_j(t) = 1, \tag{4}$$

$$M(\omega_j(t),t) = A(\omega_j(t),t), \tag{5}$$

$$\psi(\omega_j(t),t) = \theta(\omega_j(t),t) - \omega_j(t)(t-t_{p_i}). \tag{6}$$

To modify the pitch by a factor of β, the SAS system modifies the signal as follows:

$$s^M(t) = \sum_{j=1}^{\min(N(t),\lfloor\beta N(t)\rfloor)} A^M(\beta\omega_j(t),t)\cos\left[\theta^M(\beta\omega_j(t),t)\right],$$

where

$$A^M(\beta\omega_j(t),t) = a_j(t)M(\beta\omega_j(t),t), \tag{7}$$

$$\theta^M(\beta\omega_j(t),t) = \beta\omega_j(t)\left(t-t^M_{p_i}\right) + \psi(\beta\omega_j(t),t), \tag{8}$$

and the M superscript indicates a modified quantity. The modified pitch onset times are related to each other through the modified time-varying pitch period, T_0^M, as

$$t^M_{p_i} = t^M_{p_{i-1}} + T_0^M.$$

The pitch modification works as follows. The SAS system applies frequency resampling to the vocal tract components. The system is based on the assumption that the vocal tract filter is linear; therefore, the vocal tract filter outputs components of the same frequencies as those input. After pitch modification, the frequencies that enter the vocal tract filter are different from those that originally enter it. Because of this, $M(\omega,t)$ and $\psi(\omega,t)$ must be resampled along the frequency axis at the points $\beta\omega_j(t)$ for $j = 1,\ldots,\min(N(t),\lfloor\beta N(t)\rfloor)$. The SAS system modifies the amplitudes of the excitation components by frequency scaling. The

amplitude of the jth harmonic remains the same after pitch modification, but the frequency of that component is different after pitch modification. This is the reason that $a_j(t)$ is a function of j and not of $\omega_j(t)$. Finally, the SAS system modifies the phases of the excitation components by aligning all of the components on a new set of pitch marks. This excitation phase modification is expressed in the $\beta\omega_j(t)t_{p_i}^M$ term of Equation 8.

Figure 7 shows a diagram of the SAS system for pitch modification. The processing proceeds as follows.

1. Estimate the pitch, F_0, of the sampled input signal, $s(k)$, using the technique of [48]; here, k denotes the time index of the sampled signal.

2. The pitch period, $T_0 = \frac{1}{F_0}$, determines the length of a data buffer and a window, w, to be applied to the buffered data, $s_B(k)$. The window and buffer lengths are $2.5T_0$.

3. Apply the window to the buffered data, and sample the windowed data vector every 10 msec. Quatieri and McAulay found that a frame rate of 10 msec produces high-quality reconstruction.

4. Calculate the 1024-point Fast Fourier Transform (FFT) of the windowed data vector. Denote the FFT of the mth frame of data by $S(\omega, m)$. Here, ω denotes radian frequency.

5. Determine the magnitude of $S(\omega, m)$ at the center frequency of each of the 1024 bins comprising the FFT.

6. Pick the peaks of the magnitude vector, and determine the frequencies that correspond to these peaks. Denote the vector of peak magnitudes by $A(\omega, m)$ and the vector of frequencies that correspond to the peaks of the magnitude vector by $f(m)$.

7. Determine the phase of $S(\omega, m)$ at each of the frequencies given in $f(m)$; denote this phase vector by $\theta(\omega, m)$.

8. Determine $a_j(t)$, $M(\omega, t)$, and $\psi(\omega, t)$ using Equations 4, 5, and 6.

9. Modify the phases and amplitudes in order to scale the pitch by a factor of β. This involves using Equations 7 and 8. Denote the modified phase vector by $\theta^M(\omega, m)$ and the modified amplitude vector by $A^M(\omega, m)$.

10. Interpolate the amplitudes and phases on a frame-to-frame basis. This requires the matching of the frequencies of each frame with those of the next frame and uses the nearest-neighbor-based algorithm of [47].

11. Generate unit magnitude sinusoids using the frequencies and interpolated phases, then multiply each sinusoid by the corresponding interpolated amplitude.

12. Sum the sinusoids to form the pitch-modified signal, $s^M(k)$.

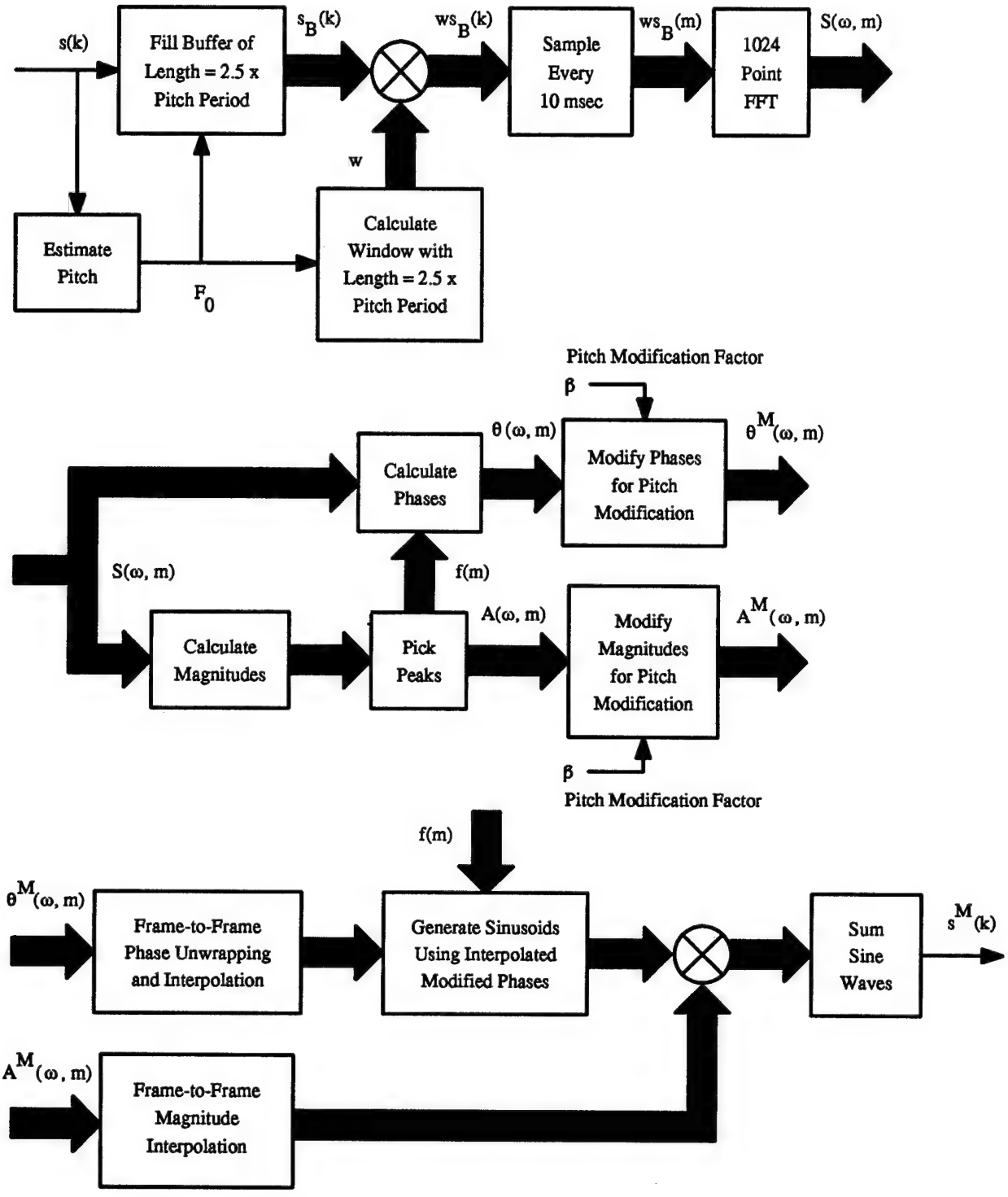


Figure 7: Pitch Modification of Speech Using the Sinusoidal Analysis/Synthesis System of Quatieri and McAulay (1992)

Quatieri and McAulay indicate in [45] that the SAS pitch modification system yields speech that is generally free of artifacts, although a certain unnatural quality is present in the unvoiced sounds due to the scaling of the unvoiced frequency components. They also indicate that some hoarseness is present for pitch modifications much greater than 20%.

The Harmonic Plus Noise Model of Laroche, Stylianou, and Moulines

As noted in the subsection concerning the multiband excitation vocoder, modeling voiced-speech segments as purely harmonic signals results in synthesized speech with a buzzy quality. For this reason, mixed-voicing models of speech such as the multiband excitation vocoder have recently been proposed. This subsection outlines a mixed-voicing model called the harmonic plus noise model (HNM) [53], which is attributed to Laroche, Stylianou, and Moulines.

The HNM system is based on a time-varying sinusoidal model for the speech signal, $s(t)$. The sinusoidal model is

$$s(t) = \sum_{n=-N(t_{p_i})}^{N(t_{p_i})} A_n(t) \exp\left(jn\omega_0\left(t_{p_i}\right)t\right) + s_S(t),$$

where $N\left(t_{p_i}\right)$ is the number of harmonics in the signal at pitch onset time t_{p_i}, $\omega_0\left(t_{p_i}\right) = 2\pi F_0$ is the radian frequency corresponding to the fundamental frequency at time t_{p_i}, $A_n(t)$ is the complex time-varying amplitude of the nth harmonic, and $s_S(t)$ is a stochastic component. The complex amplitudes are of the form:

$$A_n(t) = a_n\left(t_{p_i}\right) + \left(t - t_{p_i}\right) b_n\left(t_{p_i}\right),$$

where $a_n\left(t_{p_i}\right)$ is the complex amplitude of the nth harmonic at t_{p_i} and $b_n\left(t_{p_i}\right)$ is the complex slope of the amplitude of the nth harmonic at t_{p_i}. After some algebraic manipulation, the sinusoidal model can be written in the following form:

$$\begin{aligned} s(t) \;=\; & \sum_{n=1}^{N(t)} 2\left|a_n\left(t_{p_i}\right)\right| \cos\left(n\omega_0\left(t_{p_i}\right)t + \theta_{a_n}\left(t_{p_i}\right)\right) \\ & + \left(t - t_{p_i}\right) \sum_{n=1}^{N(t)} 2\left|b_n\left(t_{p_i}\right)\right| \cos\left(n\omega_0\left(t_{p_i}\right)t + \theta_{b_n}\left(t_{p_i}\right)\right) \\ & + a_0\left(t_{p_i}\right) + \left(t - t_{p_i}\right) b_0\left(t_{p_i}\right) + s_S(t), \end{aligned}$$

where $\theta_{a_n}(t_{p_i})$ is the phase of $a_n(t_{p_i})$ and $\theta_{b_n}(t_{p_i})$ is the phase of $b_n(t_{p_i})$. Note that $a_0(t_{p_i})$ and $b_0(t_{p_i})$ are both real-valued. Let

$$\begin{aligned}
\psi_{a_n}(t_{p_i}) &= \theta_{a_n}(t_{p_i}) + k\omega_0(t_{p_i})\,t_{p_i}, \\
\psi_{b_n}(t_{p_i}) &= \theta_{b_n}(t_{p_i}) + k\omega_0(t_{p_i})\,t_{p_i}, \\
B^C_{a_n}(t_{p_i}) &= 2\,|a_n(t_{p_i})|\cos(\psi_{a_n}(t_{p_i})), \\
B^C_{b_n}(t_{p_i}) &= 2\,|b_n(t_{p_i})|\cos(\psi_{b_n}(t_{p_i})), \\
B^S_{a_n}(t_{p_i}) &= -2\,|a_n(t_{p_i})|\sin(\psi_{a_n}(t_{p_i})), \\
B^S_{b_n}(t_{p_i}) &= -2\,|b_n(t_{p_i})|\sin(\psi_{b_n}(t_{p_i})),
\end{aligned} \tag{9}$$

then $s(t)$ can be written as

$$\begin{aligned}
s(t) = &\sum_{n=1}^{N(t)} \left\{B^C_{a_n}(t_{p_i}) + (t-t_{p_i})\,B^C_{b_n}(t_{p_i})\right\}\cos(n\omega_0(t_{p_i})(t-t_{p_i})) \\
&+\sum_{n=1}^{N(t)} \left\{B^S_{a_n}(t_{p_i}) + (t-t_{p_i})\,B^S_{b_n}(t_{p_i})\right\}\sin(n\omega_0(t_{p_i})(t-t_{p_i})) \\
&+a_0(t_{p_i}) + (t-t_{p_i})\,b_0(t_{p_i}) + s_S(t).
\end{aligned} \tag{10}$$

The HNM system modifies the pitch in the following manner. Let

$$\begin{aligned}
B^C_n(t) &= B^C_{a_n}(t_{p_i}) + (t-t_{p_i})\,B^C_{b_n}(t_{p_i}), \\
B^S_n(t) &= B^S_{a_n}(t_{p_i}) + (t-t_{p_i})\,B^S_{b_n}(t_{p_i}), \\
B_0(t) &= a_0(t_{p_i}) + (t-t_{p_i})\,b_0(t_{p_i}),
\end{aligned}$$

then $s(t)$ can be written as

$$\begin{aligned}
s(t) = &\sum_{n=1}^{N(t)} \Big\{B^C_n(t)\cos(n\omega_0(t_{p_i})(t-t_{p_i})) \\
&+B^S_n(t)\sin(n\omega_0(t_{p_i})(t-t_{p_i}))\Big\} + B_0(t) + s_S(t).
\end{aligned}$$

To modify the pitch by a factor of β, the HNM system forms the modified signal, $s^M(t)$, as

$$\begin{aligned}
s^M(t) = &\sum_{n=1}^{\min(N(t),\lfloor\beta N(t)\rfloor)} \left\{B^C_n(t)\cos\left(n\beta\omega_0(t_{p_i})\left(t-t^M_{p_i}\right)\right)\right. \\
&\left.+B^S_n(t)\sin\left(n\beta\omega_0(t_{p_i})\left(t-t^M_{p_i}\right)\right)\right\} + B_0(t) + s_S(t).
\end{aligned}$$

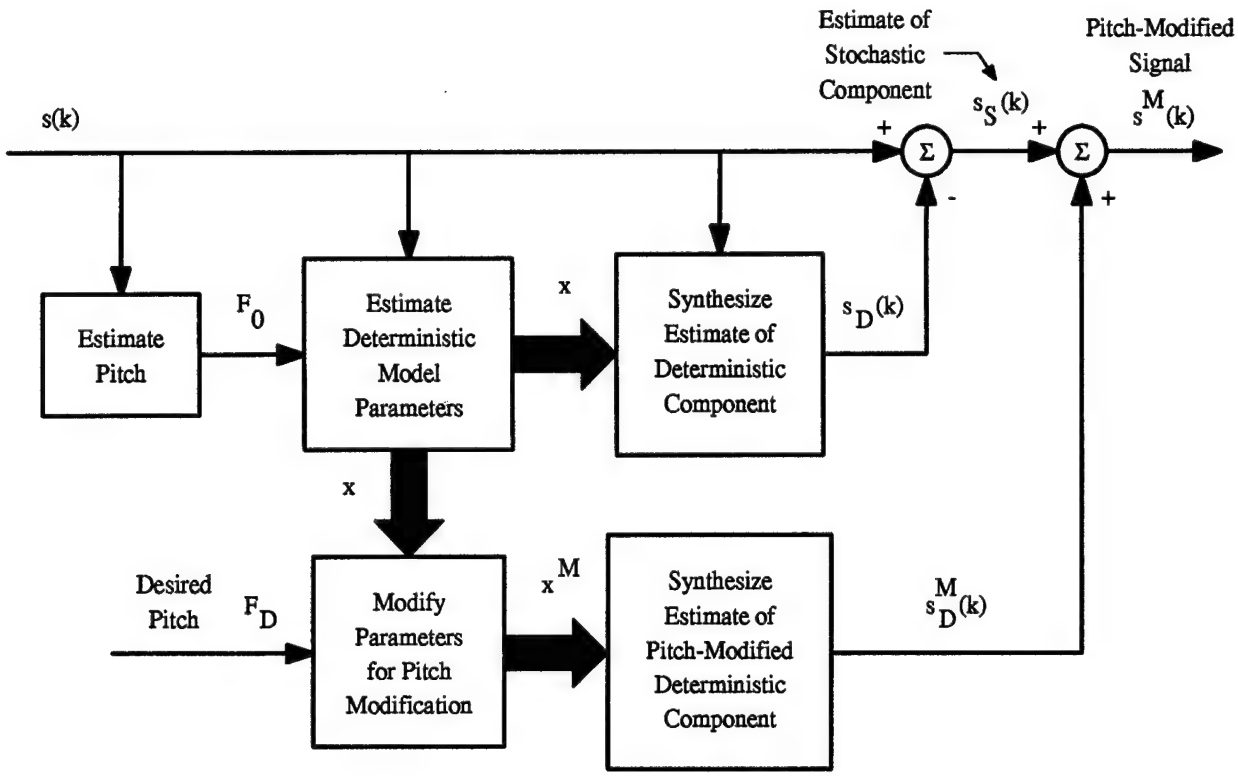


Figure 8: Pitch Modification of Speech Using the Harmonic Plus Noise Model of Laroche, Stylianou, and Moulines (1993)

Figure 8 shows a diagram of a pitch modification system based on the HNM. The system first estimates the time-varying pitch of the speech signal and the pitch onset times using, for example, the algorithm in [54]. At every pitch onset time, the system estimates the parameters of a sinusoidal model (*i.e.*, the $B_{a_n}^C(t_{p_i})$, $B_{b_n}^C(t_{p_i})$, $B_{a_n}^S(t_{p_i})$, $B_{b_n}^S(t_{p_i})$, $a_0(t_{p_i})$, and $b_0(t_{p_i})$ parameters of Equation 10) of the speech over a frame of length $2T_0$, where T_0 is the pitch period. The HNM system estimates the parameter vector, x, using a least-squares technique. The system then synthesizes the sampled estimated deterministic component, $s_D(kT_S)$, of the speech signal from x and subtracts $s_D(kT_S)$ from the original signal to yield the stochastic portion of the speech signal, $s_S(kT_S)$. If the frame is voiced, the system modifies x in order to achieve the desired pitch modification, and it uses the modified parameter vector, x^M, to synthesize an estimate of the pitch-modified deterministic component, $s_D^M(kT_S)$. If the frame is unvoiced, then the system does not modify x. Finally, the system adds $s_S(kT_S)$ and $s_D^M(kT_S)$ to form the total pitch-modified signal, $s^M(kT_S)$. Note that the $s(kT_S)$ notation is shortened to $s(k)$ in the figure.

It is instructive to compare the HNM system and the SAS. Both systems use sinusoidal models with time-varying amplitudes. The HNM system uses a time-domain least-squares technique to estimate the model parameters, while the SAS uses a frequency-domain technique based on the STFT. The SAS modifies both the stochastic and deterministic portions

of the signal, while the HNM system modifies only the deterministic portion of the signal. Quatieri and McAulay have indicated that modifying the stochastic portion in the course of modifying the pitch often leads to artifacts in the pitch-modified speech.

One of the major drawbacks of the HNM system is its large computational burden. Using a least-squares technique, the system estimates four parameters for each harmonic (see Equation 10) plus the two offset parameters ($a_0(t_{p_i})$ and $b_0(t_{p_i})$). Least-squares problems are most accurately solved using the singular value decomposition (SVD); however, the SVD is quite computationally burdensome for large numbers of parameters. Given a male speaker with an average pitch of 100 Hz and a bandwidth (BW) of 8 kHz, the HNM system estimates on the order of 318 $\left(4\left\lfloor\frac{\mathrm{BW}}{F_0}\right\rfloor - 2\right)$ parameters for each pitch period. A female speaker with an average pitch of 200 Hz and a bandwidth of 8 kHz requires approximately half the number of parameters per pitch period compared to the male speaker. However, the female speaker has twice the number of pitch periods for a given time period compared to the male speaker, so the computational burden is still high for the female speaker. Of course, lowering the speech bandwidth lowers the number of harmonics present in the speech, which in turn lowers the computational burden of the HNM system for both male and female speakers; however, lowering the bandwidth also lowers the quality of the speech.

The Pitch-Synchronous Overlap-Add Method

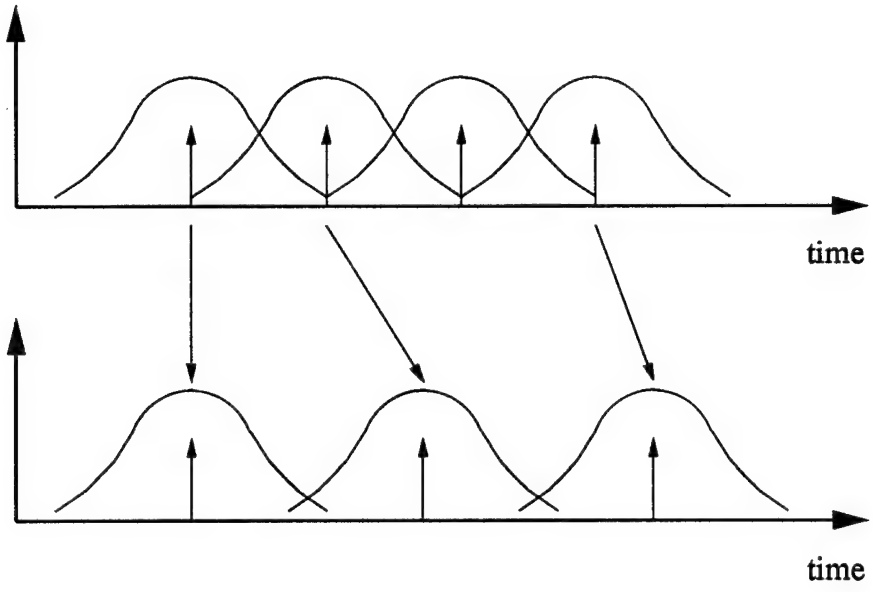
This subsection outlines the pitch-synchronous overlap-add (PSOLA) method of pitch modification [4, 55–57]. The PSOLA pitch modification technique consists of windowing segments of the original speech signal and placing these windowed segments in a (modified) output signal. In general, the speech segments have a different degree of overlap in the modified signal than they have in the original signal, and it is this change in the degree of overlap that affects the change in pitch.

The PSOLA technique builds the output signal from several windowed segments of the input signal as follows. Assume that the output signal initially consists of a zero-valued sequence, then the PSOLA process adds windowed segments of the original signal to the output one segment at a time. In performing the overlap-add (OLA) process on voiced-speech segments, the PSOLA technique makes extensive use of pitch onset times. The PSOLA technique windows the speech signal about the pitch onset times for the original signal and uses the OLA process to place the windowed segments about the pitch onset times for the output signal. The algorithm described in [54] is one popular way to estimate the pitch onset times for the original speech signal. For the output signal, one places the pitch onset times to affect the desired pitch modification.

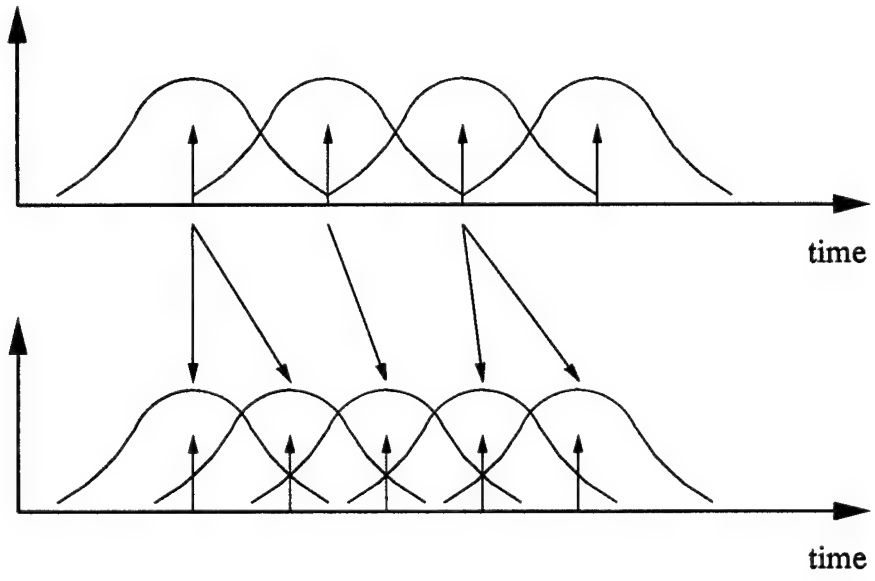
Note that in order to keep the length of the output signal approximately equal to the length of the original signal, one may need different numbers of pitch onset times for the

original and output speech signals. Figure 9 gives two examples of applying the PSOLA technique that yield different numbers of pitch onset times. Assume that the original signal has a voiced-speech segment with four pitch onset times (denoted by the small vertical arrows in Figure 9). Figure 9(a) illustrates a case where a windowed segment from the original signal is not used in the output. Figure 9(b) illustrates a case where windowed segments from the original signal are repeated in the output.

Moulines and Charpentier indicate in [4] that the PSOLA technique can introduce artifacts in the pitch-modified speech. These artifacts include tonal noises, hoarseness, and reverberation. It has been the author's experience that the PSOLA technique provides excellent quality pitch modification for pitch changes less than approximately 25%.



(a)



(b)

Figure 9: Conceptual Diagrams of the Pitch-Synchronous Overlap-Add Method for (a) Decreasing the Pitch and (b) Increasing the Pitch

MAJOR TIME-SCALE MODIFICATION TECHNIQUES

This section very briefly discusses selected time-scale modification techniques. In particular, this section outlines basic cut-and-splice methods, the synchronized overlap-add method, and wavelet-based time-scale modification methods. Note that several methods discussed in the previous section in connection with pitch modification can also be used for time-scale modification. This section does not review those techniques; instead, the reader is referred to the discussions in the previous section and in the papers.

Cut-and-Splice Methods

The general category of cut-and-splice methods encompasses several techniques; however, the basic ideas are as follows [58–62]. To reduce the duration of a speech signal to α times the original duration, remove segments of size $(1-\alpha)T$ from the signal every T seconds. To expand the duration of a speech signal to $(1+\alpha)$ times the original duration, repeat segments of size αT every T seconds. While this method is conceptually and computationally simple, it results in poor quality speech due to glitches at the segment boundaries and to the interruption of the local regularity of the speech signal [61]. In an effort to reduce these effects, Neuburg [61] propsed a system in which segments of the length of the local pitch period are discarded or repeated. Jianping [59] proposed a similar technique designed to reduce the possibility of discarding or repeating phoneme transitions. As noted in [63], these techniques often introduce a "burbling" distortion in the output speech.

The Synchronized Overlap-Add Method

A basic OLA method for time-scale modification consists of two steps. First, window the data every S_A points (*i.e.*, space the windows so that the distance between the beginning of a window and the beginning of the previous window is S_A points). Second, overlap and add the windowed segments such that the new spacing between the windowed segments is S_S points. Choose $S_S < S_A$ to compress the time scale and $S_S > S_A$ to expand the time scale.

The basic OLA method often leads to artifacts in the output speech. In an effort to reduce these artifacts, several researchers have investigated the use of the synchronized overlap-add (SOLA) method [64–69]. The SOLA technique differs from the basic OLA in that the spacing of the windows for the output speech is not S_S but $S_S + k(n)$, where $k(n)$ is a time-varying number of data points chosen to synchronize successive windowed data segments. For each segment number, n, choose $k(n)$ so that the data segments add coherently; do this by finding

the $k(n)$ that maximizes the cross correlation between adjacent windowed data segments. This process minimizes many of the glitches found in the output speech of the basic OLA method.

Wavelet-Based Methods

Time-scale modification techniques based on wavelet analysis resemble time-scale modification techniques based on the STFT. In both cases, one transforms the speech signal, modifies the parameters of the transformed signal, and computes the inverse transform to form the output signal. Let $S(p,\tau)$ be the wavelet transform of a signal, $s(t)$, then

$$S(p,\tau) = \frac{1}{\sqrt{p}} \int_{-\infty}^{\infty} g\left(\frac{t-\tau}{p}\right) s(t)\, dt,$$

where $g(\cdot)$ is the analyzing wavelet chosen by the user. The wavelet transform of $s(\alpha t)$ is related to the wavelet transform of $s(t)$ as follows:

$$s(\alpha t) \Longleftrightarrow \frac{1}{\sqrt{\alpha}} S\left(\alpha p, \alpha \tau\right).$$

Wavelet time-scale modification techniques are discussed in [70–75].

RECOMMENDATIONS AND CONCLUSIONS

This report has summarized the literature in the areas of pitch modification and time-scale modification of speech. Based on the claims made in the literature, this report recommends four techniques for further consideration—namely, the pitch-synchronous overlap-add technique, the sinusoidal analysis/synthesis system of Quatieri and McAulay, the harmonic plus noise model of Laroche, Stylianou, and Moulines, and the time-domain harmonic scaling technique. All four techniques are capable of modifying both the pitch and the time scale of speech signals. When used as pitch modification techniques, none of the four recommended techniques requires separate time-scale modification techniques. However, note that the time-domain harmonic scaling technique requires that the output signal be interpolated or decimated so that the data points are output at the same rate as the original speech signal. The pitch-synchronous overlap-add and time-domain harmonic scaling techniques have the advantage of requiring considerably less computation than do the sinusoidal analysis/synthesis system of Quatieri and McAulay and the harmonic plus noise model of Laroche, Stylianou, and Moulines. Further research remains to determine the relative performance of the recommended techniques.

REFERENCES

[1] M. Brandstein, J. Hardwick, and J. Lim, "The multi-band excitation speech coder," in *Advances in Speech Coding*, pp. 215–224, Boston, MA: Kluwer Academic, 1991. Proceedings of the IEEE Workshop on Speech Coding for Telecommunications (Vancouver, British Columbia, Canada) September 5–8, 1989.

[2] D. Griffin and J. Lim, "Multiband excitation vocoder," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 1223–1235, August 1988.

[3] P. Milenkovic, "Voice source model for continuous control of pitch period," *Journal of the Acoustical Society of America*, vol. 93, pp. 1087–1096, February 1993.

[4] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis diphones," *Speech Communication*, vol. 9, pp. 453–467, December 1990.

[5] S. Seneff, "System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, pp. 566–578, August 1982.

[6] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Paris, France), pp. 614–617, May 1982.

[7] S. Singhal and B. S. Atal, "Amplitude optimization and pitch prediction in multipulse coders," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 317–327, March 1989.

[8] M. Fratti, G. A. Mian, and G. Riccardi, "An approach to parameter reoptimization in multipulse-based coders," *IEEE Transactions on Speech and Audio Processing*, vol. 1, pp. 463–465, October 1993.

[9] B. Caspers and B. Atal, "Changing pitch and duration in LPC synthesized speech using multi-pulse excitation," *Journal of the Acoustical Society of America, Supplement 1*, vol. 73, p. S5, Spring 1983. Abstract of presentation given at the 105*th* Meeting of the Acoustical Society of America.

[10] D. Griffin and J. Lim, "A high quality 9.6 kbps speech coding system," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Tokyo, Japan), pp. 125–128, April 1986.

[11] D. Chester, "A generalized rate change filter architecture," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Minneapolis, MN), pp. III–181—184, April 1993.

[12] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1983.

[13] S. Cucchi, F. Desinan, G. Parladori, and G. Sicuranza, "DSP implementation of arbitrary sampling frequency conversion for high quality sound application," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Toronto, Ontario, Canada), pp. 3609–3612, May 1991.

[14] R. Lagadec and H. Kunz, "A universal, digital sampling frequency converter for digital audio," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Atlanta, GA), pp. 595–598, March 1981.

[15] R. Lagadec, D. Pelloni, and D. Weiss, "A 2-channel, 16-bit digital sampling frequency converter for professional digital audio," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Paris, France), pp. 93–96, May 1982.

[16] F. Ling, "Digital rate conversion with a non-rational ratio for high speed echo-cancellation modem," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Minneapolis, MN), pp. III–13–16, April 1993.

[17] S. Park, G. Hillman, and R. Robles, "A novel structure for real-time digital sample-rate converters with finite precision error analysis," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Toronto, Ontario, Canada), pp. 3613–3616, May 1991.

[18] T. Ramstad, "Sample-rate conversion by arbitrary ratios," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Paris, France), pp. 101–104, May 1982.

[19] P. P. Vaidyanathan, *Multirate Systems and Filter Banks.* Englewood Cliffs, NJ: Prentice-Hall, 1993.

[20] R. Cheung, "Real-time implementation of a 9600 bps subband coder with time-domain harmonic scaling," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Paris, France), pp. 208–211, May 1982.

[21] I. Lee and J. Gibson, "Tree coding combined with harmonic scaling of speech at 6.4 kbps," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Minneapolis, MN), pp. II–600–603, April 1993.

[22] D. Malah, "Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 121–133, April 1979. Reprinted in *Speech Enhancement*, J. S. Lim, ed., Englewood Cliffs, NJ: Prentice-Hall, 1983.

[23] D. Malah, "Combined time-domain harmonic compression and CVSD for 7.2 kbit/s transmission of speech signals," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Denver, CO), pp. 504–507, April 1980.

[24] D. Malah, R. Crochiere, and R. Cox, "Performance of transform and subband coding systems combined with harmonic scaling of speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, pp. 273–283, April 1981.

[25] J. Melsa and A. Pande, "Mediumband speech encoding using time domain harmonic scaling and adaptive residual coding," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Atlanta, GA), pp. 603–606, March 1981.

[26] J. Melsa and A. Pande, "Mediumband speech encoding using time-domain harmonic scaling and adaptive residual coding for noisy channels," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Paris, France), pp. 199–202, May 1982.

[27] J. Yuan, C. Chen, and H. Zhou, "An ADM speech coding with time domain harmonic scaling," in *Advances in Speech Coding*, pp. 367–373, Boston, MA: Kluwer Academic, 1991. Proceedings of the IEEE Workshop on Speech Coding for Telecommunications (Vancouver, British Columbia, Canada) September 5–8, 1989.

[28] M. Asi and B. Saleh, "A linear filter for time scaling of speech," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (New York, NY), pp. 79–82, April 1988.

[29] M. Asi and B. Saleh, "A linear periodically time-varying filter for time-frequency scaling of speech," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Albuquerque, NM), pp. 405–408, April 1990.

[30] C. d'Alessandro, "Time-frequency speech transformation based on an elementary waveform representation," *Speech Communication*, vol. 9, pp. 419–431, December 1990.

[31] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, pp. 235–238, June 1977.

[32] E. David, Jr., M. Schroeder, B. Logan, and A. Prestigiacomo, "Voice-excited vocoders for practical speech bandwidth reduction," *IRE Transactions on Information Theory*, vol. 8, pp. S101–S105, September 1962. Reprinted in *Speech Analysis*, R. W. Schafer and J. D. Markel, eds., New York: IEEE Press, 1979.

[33] M. Dolson, "The phase vocoder: A tutorial," *Computer Music Journal*, vol. 10, pp. 14–27, Winter 1986.

[34] H. Dudley, "The vocoder," *Bell Labs Record*, vol. 18, pp. 122–126, December 1939. Reprinted in *Speech Analysis*, R. W. Schafer and J. D. Markel, eds., New York: IEEE Press, 1979.

[35] J. Flanagan and S. Christensen, "Technique for frequency division/multiplication of speech signals," *Journal of the Acoustical Society of America*, vol. 68, pp. 1061–1068, October 1980.

[36] J. Flanagan and R. Golden, "Phase vocoder," *Bell System Technical Journal*, vol. 45, pp. 1493–1509, November 1966. Reprinted in *Speech Analysis*, R. W. Schafer and J. D. Markel, eds., New York: IEEE Press, 1979.

[37] B. Gold and C. Rader, "The channel vocoder," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, pp. 148–161, December 1967. Reprinted in *Speech Analysis*, R. W. Schafer and J. D. Markel, eds., New York: IEEE Press, 1979.

[38] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 236–243, April 1984.

[39] M. Jasiuk, V. Goncharoff, and J. Damoulakis, "Improved speech modification method," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Dallas, TX), pp. 1465–1468, April 1987.

[40] D. Malah and J. Flanagan, "Frequency scaling of speech signals by transform techniques," *Bell System Technical Journal*, vol. 60, pp. 2107–2156, November 1981.

[41] W. McGee and P. Merkley, "A real-time logarithmic-frequency phase vocoder," *Computer Music Journal*, vol. 15, pp. 20–27, Spring 1991.

[42] M. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 243–248, June 1976. Reprinted in *Speech Enhancement*, J. S. Lim, ed., Englewood Cliffs, NJ: Prentice-Hall, 1983.

[43] M. Richards, "A system for helium speech enhancement using the short-time Fourier transform," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Atlanta, GA), pp. 1097–1100, March 1981.

[44] M. Schroeder, "Vocoders: Analysis and synthesis of speech," *Proceedings of the IEEE*, vol. 54, pp. 720–734, May 1966. Reprinted in *Speech Analysis*, R. W. Schafer and J. D. Markel, eds., New York: IEEE Press, 1979.

[45] T. Quatieri and R. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Transactions on Signal Processing*, vol. 40, pp. 497–510, March 1992.

[46] R. McAulay and T. Quatieri, "Phase modelling and its application to sinusoidal transform coding," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Tokyo, Japan), pp. 1713–1715, April 1986.

[47] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 744–754, August 1986.

[48] R. McAulay and T. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal speech model," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Albuquerque, NM), pp. 249–252, April 1990.

[49] T. Quatieri and R. McAulay, "Speech transformations based on a sinusoidal representation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Tampa, FL), pp. 489–492, March 1985.

[50] T. Quatieri and R. McAulay, "Speech transformations based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 1449–1464, December 1986.

[51] T. Quatieri and R. McAulay, "Phase coherence in speech reconstruction for enhancement and coding applications," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Glasgow, Scotland), pp. 207–210, May 1989.

[52] X. Serra and J. Smith III, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, pp. 12–24, Winter 1990.

[53] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic + noise model," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Minneapolis, MN), pp. II–550–553, April 1993.

[54] B. G. Secrest and G. R. Doddington, "An integrated pitch tracking algorithm for speech systems," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Boston, MA), pp. 1352–1355, April 1983.

[55] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Tokyo, Japan), pp. 2015–2018, April 1986.

[56] K. Lent, "An efficient method for pitch shifting digitally sampled sounds," *Computer Music Journal*, vol. 13, pp. 65–71, Winter 1989.

[57] H. Valbret, E. Moulines, and J. Tubach, "Voice transformation using PSOLA technique," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (San Francisco, CA), pp. I–145–148, March 1992.

[58] G. Fairbanks, W. Everitt, and R. Jaeger, "Method for time or frequency compression-expansion of speech," *IRE Transactions on Audio and Electroacoustics*, vol. 2, pp. 7–12, January 1954. Reprinted in *Speech Enhancement*, J. S. Lim, ed., Englewood Cliffs, NJ: Prentice-Hall, 1983.

[59] P. Jianping, "Effective time-domain method for speech rate-change," *IEEE Transactions on Consumer Electronics*, vol. 34, pp. 339–346, May 1988.

[60] F. Lee, "Time compression and expansion of speech by the sampling method," *The Journal of the Audio Engineering Society*, vol. 20, pp. 738–742, November 1972.

[61] E. Neuburg, "Simple pitch-dependent algorithm for high-quality speech rate changing," *Journal of the Acoustical Society of America*, vol. 63, pp. 624–625, February 1978.

[62] R. Scott and S. Gerber, "Pitch-synchronous time-compression of speech," in *Proceedings of Conference for Speech Communications Processing*, pp. 63–65, April 1972. Reprinted in *Speech Enhancement*, J. S. Lim, ed., Englewood Cliffs, NJ: Prentice-Hall, 1983.

[63] M. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, pp. 374–390, June 1981. Reprinted in *Speech Enhancement*, J. S. Lim, ed., Englewood Cliffs, NJ: Prentice-Hall, 1983.

[64] E. Hardman, "High quality time scale modification of speech signals using fast synchronized-overlap-add algorithms," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Albuquerque, NM), pp. 409–412, April 1990.

[65] J. Makhoul and A. El-Jaroudi, "Time-scale modification in medium to low rate speech coding," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Tokyo, Japan), pp. 1705–1708, April 1986.

[66] S. Roucos and A. Wilgus, "High quality time-scale modification for speech," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Tampa, FL), pp. 493–496, March 1985.

[67] R. Suzuki and M. Misaki, "Time-scale modification of speed signals using cross-correlation functions," *IEEE Transactions on Consumer Electronics*, vol. 38, pp. 357–363, August 1992.

[68] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Minneapolis, MN), pp. II–554–557, April 1993.

[69] J. Wayman and D. Wilson, "Some improvements on the synchronized-overlap-add method of time scale modification for use in real-time speech compression and noise filtering," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 139–140, January 1988.

[70] G. Evangelista, "Pitch-synchronous wavelet representations of speech and music signals," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3313–3330, December 1993.

[71] T. Irino and H. Kawahara, "Signal reconstruction from modified wavelet transform—An application to auditory signal processing," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (San Francisco, CA), pp. I–85–88, March 1992.

[72] T. Irino and H. Kawahara, "Signal reconstruction from modified auditory wavelet transform," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3549–3554, December 1993.

[73] R. Kronland-Martinet, "The wavelet transform for analysis, synthesis, and processing of speech and music sounds," *Computer Music Journal*, vol. 12, pp. 11–20, Winter 1988.

[74] H. Ravindra, "Speech articulation rate change using recursive bandwidth scaling," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Denver, CO), pp. 352–355, April 1980.

[75] J. Youngberg, "Rate/pitch modification using the constant-Q transform," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Washington, D.C.), pp. 748–751, April 1979.

U. S. Government Printing Office 1996 750-071/00165